AUTOMATIC DETECTION OF VOICE IMPAIRMENTS DUE TO VOCAL MISUSE BY MEANS OF GAUSSIAN MIXTURE MODELS

[†]Juan I. Godino-Llorente, [†]Santiago Aguilera-Navarro, [††]Pedro Gómez-Vilda

[†] LTR (Lab. de Tecnología de Rehabilitación), Universidad Politécnica de Madrid, Ciudad Universitaria, 28041 Madrid, Spain. Ph: +34.91.336 78 32 Fax: +34.91.336 78 29 e-mail: godino@die.upm.es

[††] Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain

Abstract: **There is an increasing risk of vocal and voice diseases due to the modern way of life. It is well known that most of the vocal and voice diseases cause changes in the acoustic voice signal. These diseases have to be diagnosed and treated at an early stage. Acoustic analysis is a non-invasive technique based on digital processing of speech signal. Acoustic analysis could be a useful tool to diagnose this kind of diseases, furthermore it presents several advantages: it is a non-invasive tool, provides an objective diagnostic, moreover, it can be used for the evaluation of surgical and pharmacological treatments and rehabilitation processes. ENT clinicians use acoustic voice analysis to characterise pathological voices. In this paper, we study a well known classification approach -in speaker recognition and identification- applied to the automatic detection of voice disorders. Former and actual works demonstrate that impaired voice detection can be carried out by means of supervised neural nets: MLP (Multilayer perceptron). We have focused our task in detection of impaired voices by means of gaussian mixture models (GMMs) and parameters such mel frequency coefficients (MFFC) extracted from the windowed voice signal**.

Keywords: **Impaired Voices, MFCC, Classification, GMMs.**

I. INTRODUCTION

Acoustic analysis is a non-invasive technique based on digital processing of the speech signal. Such techniques are an effective tool for: objective support of the diagnosis; screening the vocal and voice diseases and specially their early detection; objective determination of the impairment of the vocal function; evaluation of surgical, pharmacological treatments and rehabilitation; upon obtaining an automatic analysis, the detection of some simple pathologies can be done without the presence of a specialist. Its application is not restricted to the medical area, also it can be of special interest for voice professional, as singers, speakers, etc.; evaluation of the rehabilitation.

Many algorithms to calculate acoustic parameters have been developed and, in many cases, it is demonstrated that there is a great correlation between deviations of parameters and presence or absence of impairments or pathologies [6].

This paper deals with the task of automatic detection of voice impairments by means of Gaussian Mixture Models (GMM) and Mel Frequency Cesptral Coefficients (MFCC).

II. PARAMETERIZATION

Feature extraction of speech is one of the most important issues in the field of speech technology. There are two dominant acoustic measurements of speech signal. The first is the parametric modelling approach, developed to match closely the resonant structure of the human vocal tract that produces the corresponding speech sound. It is mainly derived from Linear Predictive analysis, such as LPC and LPC-based cepstrum (LPCC) [5]. The second is the non-parametric modelling method, that is basically originated from the human auditory perception system. FFT-based mel frequency cepstral coefficients are used for this purpose. The term *mel* refers to a kind of measurement related to perceived frequency. The mapping between the real frequency scale (Hz) and the perceived frequency scales (*mels*) is approximately linear below 1KHz and logarithmic at higher frequency [5]. The bandwidth of the critical band varies according to the perceived frequency. It is about linear up to 1KHz and increases logarithmically above 1KHz. The suggested formula that models their relationship is described as follows:

$$F_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{F_{Hz}}{700}\right)$$

The advantages are that, those parameters are capable of being immune to noise and it is easy to warp frequency into a non-uniform scale, such as mel scale.

Mel frequency cepstral coefficients (MFCC) are complemented with first and second order temporal derivatives. An overview of the method is provided.

II.1. Overview of MFCC Algorithm [5]

We assume that *y[n]* denotes the input speech signal. The complete calculation process of the coefficients can be described in the next four steps as follows:

Step 1: Transform the input speech signal from time domain to frequency domain by applying short-time Fast Fourier Transform (FFT) method.

$$Y(\Omega) = \sum_{n=0}^{F-1} y[n] \cdot w[n] \cdot e^{-j 2\pi \cdot n \cdot \frac{m}{F}}$$

where *m = 0,1,2…F-1; F* is frame size, which is generally equal to the power of 2; *w[n]* is the Hanning window function, which is based on the fact that the signal can be regarded as stationary and uninfluenced by the others within a short period of time, i.e. the frame size.

Step 2: Find the energy spectrum of each frame.

$$X(\Omega) = |Y(\Omega)|^2$$

Step 3: Calculate the energy in each mel window.

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25OCT2001	N/A	-

Title and Subtitle Automatic Detection of Voice Impairments due to Vocal Misuse by Means of Gaussian Mixture Models	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) LTR (Lab. de Tecnología de Rehabilitación), Universidad Politécnica de Madrid, Ciudad Universitaria, 28041 Madrid, Spain.	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

$$S_k = \sum_{j=0}^{\frac{k}{2}-1} W_k(j) \cdot X(j)$$

where $1 \le k \le M$; M is the number of the mel windows in mel scale, which generally ranges from 20 to 24. $W_k(j)$: the triangular weighted function is associated with the k^{th} mel window in mel scale.

Step 4: Proceeding with logarithm and cosine transforms, we can figure out the mel frequency cepstral coefficients:

$$mc_m = \sum_{k=1}^{M} \log(S_k) \cos\left[n \cdot (k - 0.5) \frac{\pi}{M} \right]$$

where $1 \le n \le L$; L is the desired order of MFCC.

II.2. Temporal Derivative [5]

An improved representation can be obtained by extending the analysis to include information about the temporal parameters derivative. Both first and second derivatives have been used. To introduce temporal order into the parameter representation, we denote the m^{th} coefficient at time t by $c_m(t)$:

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu \cdot \sum_{k=-K}^{K} k \cdot c_m(t+k)$$

where μ is an appropriate normalization constant and $(2K+1)$ is the number of frames over which computation is performed.

For each frame t, the results of the analysis is a vector of Q coefficients, and appended to it two Q length vectors more of first and second time derivatives; that is:

$$o(t) = (c_1(t), c_2(t), ..., c_Q(t),$$
$$\Delta c_1(t), \Delta c_2(t), ..., \Delta c_Q(t),$$
$$\Delta\Delta c_1(t), \Delta\Delta c_2(t), ..., \Delta\Delta c_Q(t))$$

where $o(t)$ is a vector with $3 \cdot Q$ components.

III. GAUSSIAN MIXTURE MODEL (GMM) [2] [3]

Let $x \in \Re^n$ be a random vector that has an arbitrary distribution. The distribution density of x is modelled as a Gaussian Mixture Density, a mixture of Q component densities, given by:

$$p(x/\lambda) = \sum_{i=1}^{Q} c_i \cdot p_i(x), \quad \sum_{i=1}^{Q} c_i = 1, \quad c_i \ge 0$$

where $p_i(x)$, $i=1,...,Q$ are the component densities and c_i, $i=1,...,Q$ are the component weights. Each component density is a n-variate Gaussian function of the form:

$$p_i(x) = \frac{1}{(2\pi)^{n/2} |C_i|^{1/2}} \exp\left[-\frac{1}{2} (x-\mu)^T C_i^{-1} (x-\mu) \right]$$

with μ_i the $n \times 1$ mean vector and C_i the $n \times n$ covariance matrix.

The complete Gaussian mixture density is parameterised by the mean vectors, covariance matrices and mixture weights from all component densities, and these parameters are represented by the notation

$$\lambda = \{c_i, \mu_i, C_i\} \qquad i = 1, ..., Q$$

As the Gaussian components are acting together to model the overall pdf, full covariance matices are not necessary. The linear combination of diagonal covariance Gaussians is capable of modelling the correlation between feature elements. With this asumption, the parameters of the mixture can be represented by:

$$\lambda = \{c_i, \mu_i, \sigma_i\} \qquad i = 1, ..., Q$$

where σ_i is a $n \times 1$ vector representing the diagonal of the covariance matrix.

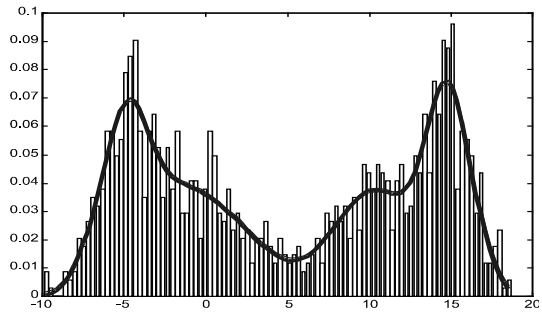

Figure 1: Histogram of a single cesptral coefficient and its aproximmation by means of a gaussian mixture.

There is a principal motivation for using the GMM as a representation of the acoustic space: it is the empirical observation that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions.

III.1. EM algorithm [3]

To model the distribution of the acoustic space, parameters λ of the GMM may be estimated using MFCC coefficients. There are several techniques available for estimating the parameters of the GMM. The most popular method is maximum likelihood (ML) estimation. ML parameter estimates can be obtained iteratively using the well-known expectation and maximization (EM) algorithm.

The goal of the EM algorithm is to estimate the paramenters of the GMM, λ which best matches the distribution of the training samples. For a sequence of T training vectors $X = \{\vec{x}_1, \vec{x}_3, ..., \vec{x}_T\}$ the GMM likelihood can be written as:

$$p(X/\lambda) = \prod_{t=1}^{T} p(\vec{x}_t / \lambda)$$

But, this expression is a non-linear function of the parameters λ, and direct maximization is not possible. EM algorithm allows us to estimate the model parameters. The idea is beginning with an initial new model λ', such that $p(X/\lambda') \ge p(X/\lambda)$. The initial model becomes the new model for the next iteration, and so on. On each EM iteration the reestimation formulas used to guarantee a convergence of the model are:

$$c_i = \frac{1}{T} \sum_{t=i}^{T} p(i/x_t, \lambda) \qquad \mu_i = \frac{\sum_{t=i}^{T} p(i/x_t, \lambda) x_t}{\sum_{t=i}^{T} p(i/x_t, \lambda)}$$

$$\sigma_i^2 = \frac{\sum_{t=i}^{T} p(i/x_t,\lambda)\cdot x_t^2}{\sum_{t=i}^{T} p(i/x_t,\lambda)} - \mu_i^2$$

The a posteriori probability for acoustic class m is given by the expression:

$$p(m/x_t,\lambda) = \frac{c_m p_m(x_t)}{\sum_{t=1}^{T} c_k p_k(x_t)}$$

IV. DATABASE

The company Kay Elemetrics recorded to CD-ROM a database of 1,400 voice samples from approximately 700 subjects. The Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs originally developed this database [1].

The acoustic samples are sustained phonation and running speech samples from patients with normal voices and a wide variety of organic, neurological, traumatic, and psychogenic voice disorders.

The speech samples were collected in a controlled environment and sampled with a 25 kHz sampling rate and 16 bit of resolution.

The database contains sustained phonation of vowels (3-4 sec. long) and running speech samples of the *rainbow passage*.

Data have been divided into two subsets: the first subset has been used for training (70%), the second (30%), to simulate and validate results.

V. METHODOLOGY

In the introduction we pointed out that the focus of this research was the automatic detection of voice disorders. Figure 2 shows a block diagram explaining how the pre-processing front-end works. Firstly, speech is filtered to avoid aliasing. Secondly, the signal is converted into a sequence of samples by the A/D converter; later, speech is enframed using 1024 samples Hanning windows. At this point the signal is preprocessed using a preemphasis filter. The next module is an endpoint detector in order to avoid unvoiced segments or silences. It is important to remove silence/noise frames from both training and testing signals to avoid modeling the environment. After the endpoint detection module, frames are parameterised in order to reduce dimesionality: MFCC parameters are calculated. Each register is quantified using a n-dimensional vector composed by the MFCC coefficients plus their first and second derivative. The detector have been tested with and without pre-emphasis filtering.

Last module (GMM) is related with the Gaussian mixture based classifier. The learning algorithm that has been used is EM (expectarion maximization) [3]. Such architecture is widely used in speaker verification and recognition [2] [3]. The number of gaussians we have used is 32, 64 and 128.

Training were carried out over 20 epochs. At that point the aproximation error raises below 1%. Sum of mean squared error is controlled as parameter to stop training. Model initialization is carried out by means of k-means clustering algorithm. Covariance matrices used are diagonal.

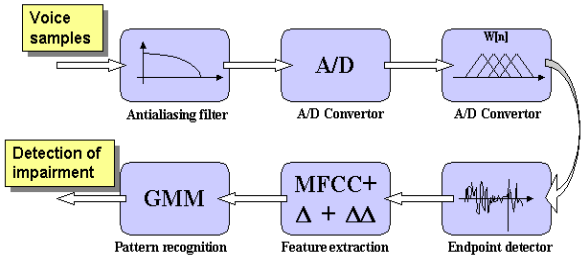

Figure 2: Preprocessing front-end

The final selected number of voice samples from the database was 106 (53 normal and 53 pathological voices).

As the pre-processing front-end divides speech signal into overlapped frames, we will dispose of one input vector per frame for training the classifier. The total amount of vectors used to train the system is around 25.000, each corresponding to a framed window. Nearby 50% correspond to normal voices, and 50% remaining to pathological ones.

VI. PERFORMANCE MEASUREMENTS

It is applied the likelihood ratio test to an utterance to determine if the voice register is normal or not. For an utterance $X = \{\vec{x}_1, \vec{x}_3, ..., \vec{x}_T\}$ and a model λ_c, the likelihood ratio is:

$$\frac{\Pr(X \quad is \quad normal)}{\Pr(X \quad is \quad patho\log ical)} = \frac{\Pr(\lambda_C / X)}{\Pr(\lambda_{\overline{C}} / X)}$$

Appliying Bayes' rule and discarding the constant prior probabilities, the likelihood in the *log* domain is:

$$\Lambda(X) = \log\left[p(X/\lambda_C)\right] - \log\left[p(X/\lambda_{\overline{C}})\right]$$

The likelihood ratio is compared with a threshold θ and the voice is said to be pathological if $\Lambda(X) > \theta$ and normal if $\Lambda(X) < \theta$. The decision threshold is then set to adjust the trade off between rejecting pathological voices (false rejection) or accepting normal voices (false aceptance error).

The computation of the likelihood ratio is as follows: the likelihood of the utterance given the normal voice model is computed as:

$$\log\left[p(X/\lambda_C)\right] = \frac{1}{T}\sum_{t=1}^{T} \log\left[p(x_t/\lambda_C)\right]$$

The likelihood of the utterance given the pathological vocice model is:

$$\log\left[p(X/\lambda_{\overline{C}})\right] = \frac{1}{T}\sum_{t=1}^{T} \log\left[p(x_t/\lambda_{\overline{C}})\right]$$

A block diagram of the detector is shown in Figure 3.

For a given input test utterance X the choice is between *s* (X is a normal voice register) and *n* (X is a pathological voice register). Possible decisions are: *S,* the register is classified as normal; *N,* the register is classified as pathological.

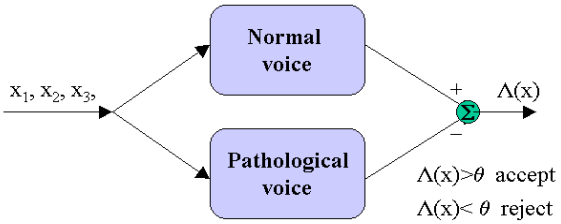

Figure 3: Scheme of the detector. From each frame, the likelihood is calculated for both models.

Outputs:

- *P(S/s):* probability of correct detection (CD). Detector found an event (presence of pathology) when one was present.
- *P(S/n):* probability of false detection (FD). Detector found no event (normal voice) when one was present (pathological voice)
- *P(N/n):* probability of correct rejection (CR). Detector found no event when indeed none was present.
- *P(N/s):* probability of false rejection (FR). the Detector found an event when none was present.

These probabilities satisfy next conditions:

$$p(S|s)+p(N|s)=1 \qquad p(S|n)+p(N|n)=1$$

VII. RESULTS

Figure 4 shows a typical false aceptance (clear line) and false rejection (dark line) plot vs. threshold. ERR is the point where both lines intersect.

Presence or absence of impairment decision is carried out taking into account the number of vectors belonging to each class. Table 1 shows frame accuracy resutls with 32, 64 and 128 gaussians, with and without preemphasis.

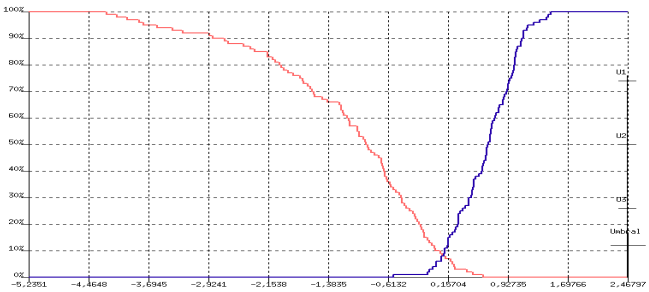

Figure 4: False aceptance (right) and false rejection (left). Both lines cross in the equal error rate (ERR) point.

Best results are obtained using a *64 or 128 GMM* mixture, trained during 20 epochs with *15 or 17 MFCC plus ΔMFCC+ΔΔMFCC+Energy+ΔEnergy+ΔΔEnergy* calculated from sustained vowels. In such case, and using this database, ERR is around 0,1%. Results shows that preemphasis does not improve detector accuracy.

MFCC prmph	11	12	13	14	15	16	17	18	19	20
GMM32	41,5	22,3	32,9	44,4	41,7	35,3	49,4	35,4	41,2	12,1
GMM64	40,8	40,5	41	44,5	22	43,2	50	50	36,6	37,4
GMM128	40,4	43,9	41,5	37,8	43,3	47,1	46,9	41,3	40,6	39,1

MFCC	11	12	13	14	15	16	17	18	19	20
GMM32	29,5	18,3	19,7	45	3,3	43,4	22,9	17,5	20,3	26,3
GMM64	3,7	0,5	1,7	11,3	0,1	11,3	0,2	1,3	0,29	0,8
GMM128	3,1	2,5	0,7	0,18	0,11	0,45	0,02	0,15	0,8	0,9

Table 1: ERR vs. parameters (frame accuracy)

EVENT			
DECISION		ABSENT	PRESENT
	ABSENT	CR≈100%	FR≈0%
	PRESENT	FD≈0%	CD≈100%

Table 2: Performance matrix of the detector (register accuracy)

Using this scheme, all registers were rightly classified.

VII. CONCLUSIONS & FUTURE WORK

GMM technology seems to be a promisable tool for the automatic detection of voice disorders. In any case, a notable attention should be paid, because the database stores a collection of very significant medical cases. Conclusions have to be tested with a larger database. Results seems to be better with that using MLP [7][8]. In order to compare both techniques confidence measurments have to be done. Preemphasis does not improve performance.

Due to the fact that it seems to be easy to detect voice disorders, the next step will be to classify a set of pathologies. For this purpose, a two stepped scheme may be used: the first step will deal with the detection of the presence of voice disorders; once the presence is confirmed, in the second step it will be guessed what kind of disorder it is.

Anycase, a wider database of pathological voices is needed. What's more it has to be tested using running speech.

REFERENCES

[1] "Disordered Voice Database", Version 1.03, Kay Elemetrics Corp, 1994

[2] "Speaker identification and verification using Gaussian mixture seaker models" D. A. Reynolds. Speech Communication, 1995, pp 91-108

[3] "Robust text-independent speaker identification using gaussian mixture seaker models" D. A. Reynolds, R.C. Rose. IEEE Transactions on speech & audio processing , Vol. 3 No. 1, 1995, pp 72-83

[4] "Neural networks for pattern recognition" C.M. Bishop. Oxford University Press, 1995.

[5] "Fundamentals of speech recognition" L. Rabiner. Prentice Hall. 1993.

[6] "Clinical measurement of speech and voice" R.J. Baken. Taylor & Francis. 1993.

[7] JI. Godino-Llorente et al. "On the selection of meaningful speech parameters used by a pathologic/non pathologic voice register classifier". Proceedings of Eurospeech'99, Budapest, Hungary, 1999, pp 563-566

[8] JI. Godino-Llorente et al. "LPC, LPCC and MFCC parameterization applied to the detection of voice impairment". Proceedings of ICSLP'00, Beijing, China, 2000.